JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 111, D23S30, doi:10.1029/2006JD007124, 2006

Ensemble-based ozone forecasts: Skill and economic value

Mariusz Pagowski[1,2] and Georg A. Grell[1,3]

Received 25 January 2006; revised 17 May 2006; accepted 10 July 2006; published 16 November 2006.

[1] In the summer of 2004, seven air quality models provided forecasts of surface ozone concentrations over the eastern United States and southern Canada. Accuracy of these forecasts can be assessed against hourly ozone measurements at over 350 locations. The ensemble of the air quality models is used to issue deterministic and probabilistic forecasts of maximum daily 8-hour and 1-hour averaged ozone concentrations. For completeness, a short summary on performance of deterministic forecasts for this ensemble of models, obtained alternatively by averaging model concentrations or by using dynamic linear regression as described by Pagowski et al. (2006), is given on the basis of this work. In parallel, the skill of probabilistic forecasts is discussed. To remove the bias, the probabilistic forecasts are calibrated. The economic value of forecasts, which is calculated using Richardson's cost-loss decision model, is evaluated for both deterministic and probabilistic cases. It is shown that deterministic forecasts obtained with the ensemble of models provide a greater benefit to decision makers than forecasts issued with individual models. Probabilistic forecasts demonstrate similar advantages over the deterministic forecasts.

Citation: Pagowski, M., and G. A. Grell (2006), Ensemble-based ozone forecasts: Skill and economic value, *J. Geophys. Res.*, *111*, D23S30, doi:10.1029/2006JD007124.

1. Introduction

[2] Predictions of atmospheric models are inherently uncertain as they depend on the initial states and forcings [e.g., *Kalnay*, 2003]. This fact provides a justification for ensemble forecasting. Today, weather prediction with ensembles is a common practice at meteorological centers around the world. The rationale for using ensembles in air quality is as strong as in weather prediction, taking into account the complexity of chemical processes, their dependence on the atmospheric state, uncertainty in equations which attempt to describe chemical reactions, and questionable quality of emission inventories. Nevertheless, despite the compelling evidence for sensitivity of chemical models to meteorology, parameterizations, and emission inventories [*Russell and Dennis*, 2000], ensemble modeling in air quality is rare. Previously, *Vautard et al.* [2001], *Delle Monache and Stull* [2003], *Pagowski et al.* [2005, 2006], *McKeen et al.* [2005], and *Delle Monache et al.* [2006a, 2006b] used different techniques to demonstrate the advantage of deterministic ensemble forecasts compared with forecasts provided by individual models.

[3] Skill of weather and air quality forecasts is assessed using a variety of measures. Commonly used skill scores for deterministic forecasts include hits and misses ratios, probability of detection, threat score, and bias ratio [e.g., *Wilks*, 1995]. For probabilistic forecasts, the Brier score [*Brier*, 1950] and relative operating characteristic (ROC) [*Mason*, 1982] are frequently used. However, utility of forecasts is most appropriately judged by the benefits they provide to users. The economic value of forecasts was first discussed by *Thompson and Brier* [1955] using a two-state (occurrence/no occurrence) two-action (action/no action) cost-loss model. *Kernan* [1975] applied this model to maximize benefits of air quality forecasts in California. Multiple authors have since investigated the relationship between skill and economic value of weather forecasts, concluding that this relationship is complex and that skill alone may provide misguided expectation of the practical value of forecasts [e.g., *Katz and Murphy*, 1997; *Richardson*, 2000].

[4] Most recently applications of cost-loss models to assess a potential economic value of weather forecasts were presented by *Mylne* [1999], *Richardson* [2000, 2003], and *Wilks* [2001]. *Richardson* [2000] studied the economic value of forecasts of temperature and precipitation at the European Centre for Medium-Range Weather Forecasts (ECMWF) with a static cost-loss model. The author demonstrated that the operational probabilistic forecasts issued with the ensemble prediction system (51 ensembles at T159 horizontal resolution) provided generally greater benefit than the high-resolution deterministic forecasts (at T319 horizontal resolution) and suggested that the benefit of the ensemble system is equivalent to many years of develop-

[1]Global Systems Division, Earth System Research Laboratory, NOAA, Boulder, Colorado, USA.
[2]Also at Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado, USA.
[3]Also at Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA.

0148-0227/06/2006JD007124$09.00

ment of the deterministic model and assimilation system. With the application of the same economic model to evaluate the economic value of forecasts of 500 hPa geopotential height at National Centers for Environmental Prediction (NCEP), *Zhu et al.* [2002] showed that a 14-member ensemble (at T62 horizontal resolution) provides a greater benefit than the high-resolution deterministic model (at T164 horizontal resolution).

[5] In this paper we assess skill and a potential economic value of probabilistic forecasts of surface ozone from an ensemble of air quality models compared to deterministic forecasts. The deterministic forecasts are obtained for this ensemble by averaging models results and using dynamic linear regression [*Pagowski et al.*, 2005] and also include the forecasts of individual models.

[6] The data come from seven air quality models which participated in the International Consortium for Atmospheric Research on Transport and Transformation/New England Air Quality Study (ICARTT/NEAQS) conducted over the eastern United States and southern Canada in the summer of 2004 and provided daily forecasts of surface ozone. For verification, hourly averaged surface ozone concentrations at over 350 locations are available from the Aerometric Information Retrieval Now (AIRNow) database.

[7] The observations and models are presented in section 2. Deterministic forecasts are described in section 3. Section 4 gives details on verification and calibration of the probabilistic forecasts. In section 5, the economic value of surface ozone forecasts is assessed using a simple cost-loss *Richardson*'s [2000, 2003] model. Conclusions are drawn in the final section.

2. Observations and Models

2.1. Observations

[8] Since only a short description of ozone observations is given here, the reader is referred to *McKeen et al.* [2005] for more detailed information. Hourly surface ozone measurements available at over 350 sites located within the modeling domain from 0000 UTC 6 July to 0000 UTC 30 August 2004 (56 days) were used to calculate maximum daily 8-hour and 1-hour averaged ozone concentrations. Figure 1 shows locations of the sites, the AIRNow site classification, and outline of the domain of model overlap. The prevailing cold and rainy weather over the area of consideration during this period lead to few high ozone episodes. Out of 16480 observations, only 87 exceedances of an EPA-mandated 85 ppbv threshold for the maximum daily 8-hour averaged ozone concentration, and no exceedances of a 125 ppbv threshold for the maximum daily 1-hour averaged ozone concentration were recorded. The small number of high ozone observations limits our ability to fully assess performance of the ensemble with respect to these EPA-mandated thresholds.

2.2. Models

[9] The seven air quality models participating in the ICARTT/NEAQS field study that constituted an ensemble include AURAMS [*Moran et al.*, 1998], MAQSIP [*McHenry and Coats*, 2003] (available at two horizontal resolutions), CHRONOS [*Pudykiewicz et al.*, 1997], CMAQ [*Byun and Ching*, 1999], STEM-2K3 [*Carmichael et al.*, 2003], and WRF/Chem [*Grell et al.*, 2005]. These models used different meteorological drivers, chemical mechanisms, emission inventories, and methods for their processing, and were executed at varying horizontal and vertical resolutions. In this study, 24-hour forecasts of the models issued at 0000 UTC (0600 UTC for CMAQ) were used. The domain of model overlap is shown in Figure 1.

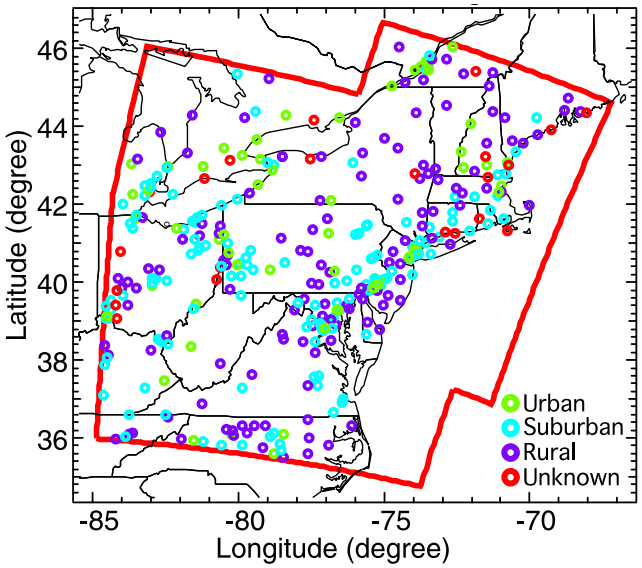

Figure 1. Locations of the measurement sites, the AIRNow site classification, and outline of the domain of model overlap.

[10] For verification, model values were matched with the site measurements located within the grid cell. For consistency with observations, results from models output at the top of the hour were averaged temporally. *McKeen et al.* [2005] provide further details on the models and processing of results for verification.

3. Deterministic Forecasts

[11] Commonly, ensemble forecasts are used to provide probabilistic distribution of the future scenario (probabilistic forecasts). They can also be used to provide best estimates of the future state of the atmosphere (deterministic forecasts). Deterministic ensemble forecasts of ozone concentrations during the ICARTT/NEAQS field study were extensively discussed by *McKeen et al.* [2005], *Pagowski et al.* [2005, 2006] and *Delle Monache et al.* [2006a, 2006b] and only a short summary is given here for completeness.

[12] Verification of ozone forecasts revealed the presence of positive bias which was most evident at lower concentration thresholds. It was demonstrated that the different bias removal techniques had a positive effect on root mean square error and skill scores such as the probability of detection, hits and misses ratios, and equitable threat scores, but did not improve correlation. A correlation of the ensemble average was higher than that of any individual model. The maximum benefits of bias removal techniques were noted for lower concentration thresholds. The reason for the limited success of bias removal techniques at the higher concentration thresholds is a result of relative rarity

of high ozone during the field experiment and insufficient sensitivity of these methods to the different magnitude of biases at different thresholds.

[13] In the current study, the bias removal technique employed by *Pagowski et al.* [2006] is applied. This method devised by *West and Harrison* [1989] is referred to as dynamic linear regression (DLR) and compared to static linear regression is more responsive to the variability of the process. With this method, a set of weights which vary temporally and spatially is devised for each ensemble member. A deterministic forecast is obtained by multiplying forecasts of ensemble members by their specific weights. Performance of this method was comparable to the method used by *Pagowski et al.* [2005] based on least square error minimization using ensemble forecasts from the previous day. Further details on the algorithm are given by *Pagowski et al.* [2006] and *West and Harrison* [1989, pp. 60–71, 117–121].

4. Probabilistic Forecasts

4.1. Skill

[14] Given the inherent uncertainty of predictions of chemical models, probabilistic forecasts have a natural advantage over deterministic forecasts in that they provide an estimate of likelihood of an occurrence of an event on a scale from 0 to 1 rather than a categorical statement (“maybe” versus “yes” or “no”). Thus probabilistic forecasts intuitively contain more information. Forecast probability of an ensemble is calculated as a fraction of forecasts predicting a binary event among all the considered forecasts. A binary event has only two possible outcomes: either it happens or not. For a continuous predictant, the occurrence of an event such as surface ozone concentration is assessed, if a predicted value is greater than a certain threshold.

[15] The half-Brier score (*BS* hereinafter) [*Wilks*, 1995, pp. 259–263] is frequently used in assessing skill of probabilistic forecasts and is defined as a mean square error of the forecast probability

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2, \tag{1}$$

where N is the number of forecasts, p_i is the forecast probability, and o_i is the observation which takes the value of 1 when the event occurs and 0 if it does not. *BS* is equal to 0/1 when all the forecasts accurately/inaccurately predict occurrence of an event or series of events. The half-Brier score is commonly called Brier score or probability score despite the fact that it differs from the measure originally devised by *Brier* [1950]. *Murphy* [1973] expressed *BS* as a sum of three terms:

$$BS = \sum_{k=1}^{K}\frac{n_k}{N}(p_k - \overline{o}_k)^2 - \sum_{k=1}^{K}\frac{n_k}{N}(\overline{o}_k - \overline{o})^2 + \overline{o}(1-\overline{o}), \tag{2}$$

where N forecasts were divided into K categories, each comprising n_k cases of forecast probability p_k; $\overline{o}_k$ is an observed frequency when forecast probability is p_k (i.e., conditional frequency of events given that they are forecast), and $\overline{o}$ is the observed frequency of events in the sample. It is apparent that in the case of an ensemble with m members, the maximum number of categories is $m + 1$ (forecast probability varies from 0 to m/m). The first term in the *BS* decomposition called reliability reflects on the bias of forecasts (the smaller the better). The second term, resolution, indicates the ability of an ensemble to distinguish between different categories (via conditional probability of observations, the larger the better). The third term, uncertainty, depends on the variance of observations and measures the difficulty of forecasting during the considered period. Uncertainty is thus independent of the quality of forecasts.

[16] In Figure 2, *BS* and its components for three different thresholds of the maximum daily 8-hour and 1-hour averaged ozone concentrations (50 ppbv, 70 ppbv, and 85 ppbv) and using eight forecast probability categories are plotted for the ensemble of air quality models executed in the summer of 2004. It should be noted that the values of *BS* for the different thresholds cannot be directly compared as a consequence of varying difficulty of predictions for these thresholds (because of varying uncertainty terms). For example, ozone concentrations are almost as likely to be above as below 50 ppbv and are thus most difficult to predict. Measurements above the higher thresholds are much less common and just with the knowledge of this climatology, it can be predicted that ozone concentrations will generally fall below these thresholds. Large values of reliability accompanied by small values of resolution, especially evident for 70 and 85 ppbv thresholds, are signs of shortcomings of the ensemble.

[17] The reliability diagram, which illustrates correspondence between the forecasts probabilities (x axis) and observed frequencies (y axis) in the reliability term, provides further insight into the deficiencies of the ensemble. In the ideal case, the reliability curve overlays the 45° line. Frequency of forecasts within the categories (their sharpness) is plotted in the insets along with the average observed frequency (unconditional, i.e., climatology of the sample). It is desirable that frequencies of forecasts fall into 0 or 1 bins as this indicates that an ensemble has a skill in predicting presence of absence of events. In Figure 3 reliability diagrams are plotted for the ensemble forecasts of maximum daily 8-hour and 1-hour averaged ozone concentrations and the above thresholds. It can be noted in Figure 3 that forecast probabilities are higher than observed frequencies and thus the ensemble shows marked overforecasting bias for all the thresholds. Flatness of the curves for the highest threshold, most prominent in the case of maximum daily 8-hour averages, is a manifestation of a poor resolution. Uneven trend of the same curve indicates a small sample size. In the insets, sharpness of the forecasts of concentrations larger than 50 ppbv is skewed toward higher probabilities. Since observed frequencies of exceedances of 85 ppb threshold for the 8-hour and 1-hour averages are equal to 0.384 and 0.522, respectively, the skewness is another indication of the overforecasting bias for this threshold (ideally the diagram should be U-shaped). Nevertheless, by assigning maximum probabilities toward 0 and 1 bins, the ensemble distinguishes between events when ozone concentration thresholds are exceeded or not and just shows skill compared to climatological forecasts. For the

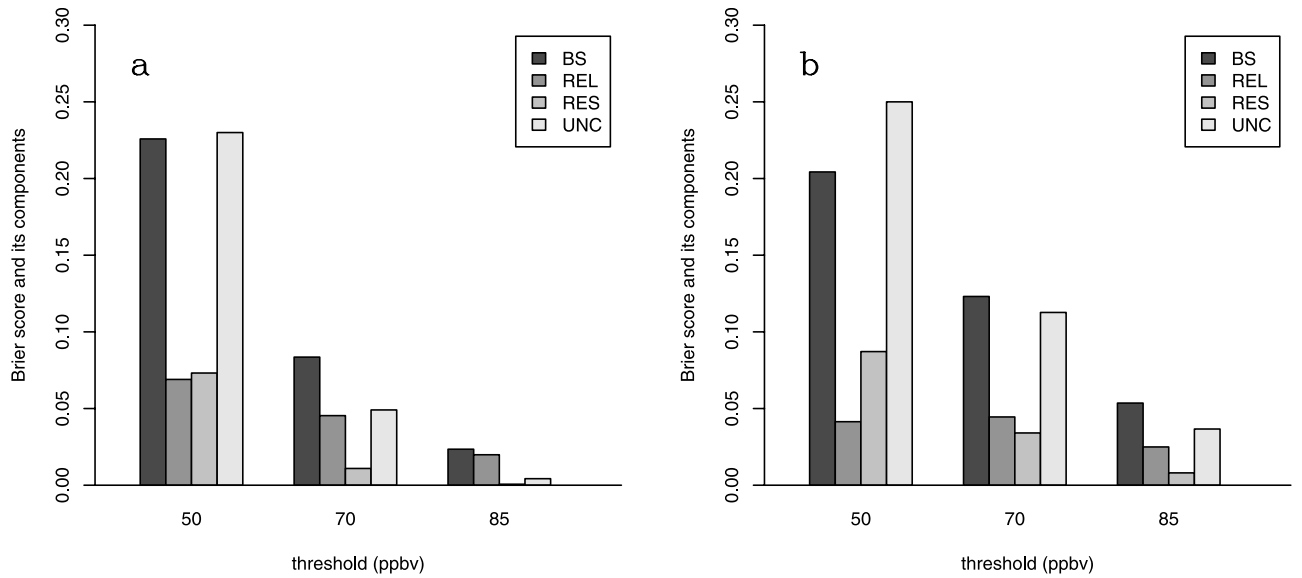

Figure 2. Brier score and its components of forecasts of maximum daily (a) 8-hour and (b) 1-hour averaged surface ozone concentration over 50 ppbv, 70 ppbv, and 85 ppbv thresholds.

latter seasonal average ozone concentration is assigned daily and such forecasts have no sharpness.

4.2. Calibration

[18] It is a common practice in weather forecasting, called calibration, to match forecast probabilities with the observed frequencies for the past events where the same ensembles are executed over extended periods of time [e.g., *Wilks*, 1995]. This method is, however, not useful in the current case since the ensemble of air quality models was assembled ad hoc for the ICARTT/NEAQS field study, and its past performance is unknown. Bias removal does not typically improve reliability of an ensemble as much as calibration based on the past performance of the ensemble [*Atger*, 2003], but remains the only available option here. For this purpose, dynamic linear regression (DLR) in a form described shortly in the previous section and fully by *Pagowski et al.* [2006] is applied. The only difference between the present and former application of the algorithm is that now two regression coefficients are sought for each model separately while previously, weights were sought for ensemble members to produce a single deterministic forecast.

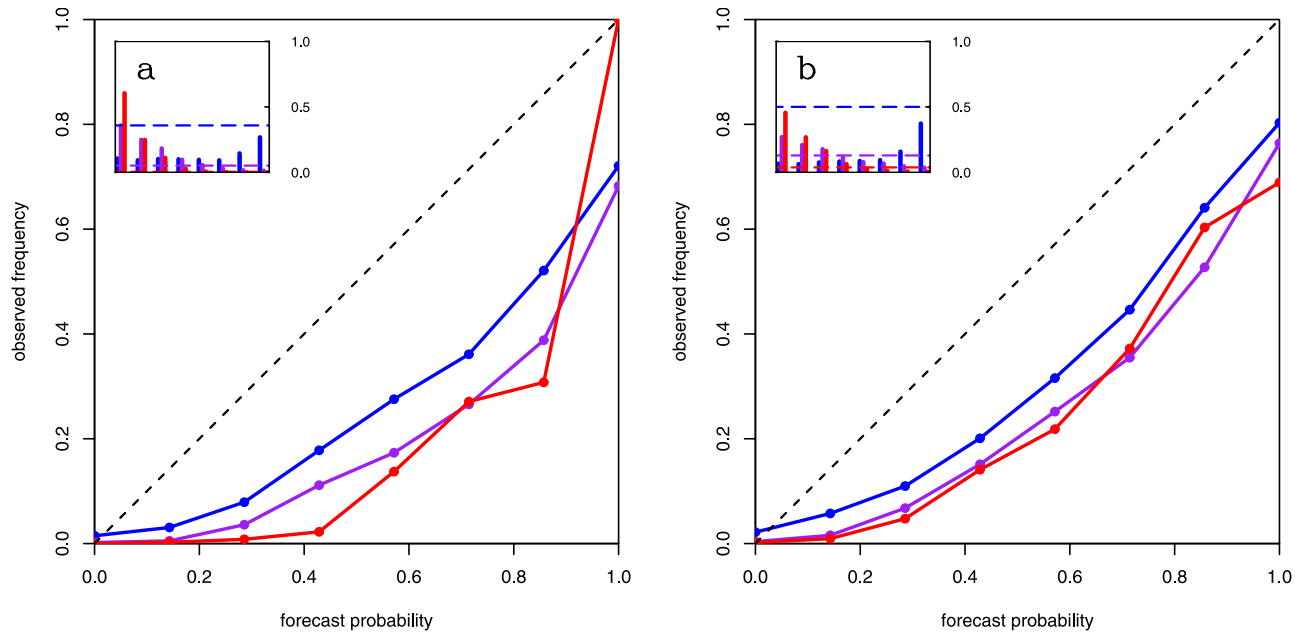

Figure 3. Reliability diagrams for maximum daily (a) 8-hour and (b) 1-hour averaged surface ozone concentration over 50 ppbv (blue), 70 ppbv (purple), and 85 ppbv (red) thresholds. Inset: Forecasts probability and observed frequency (dashed lines).

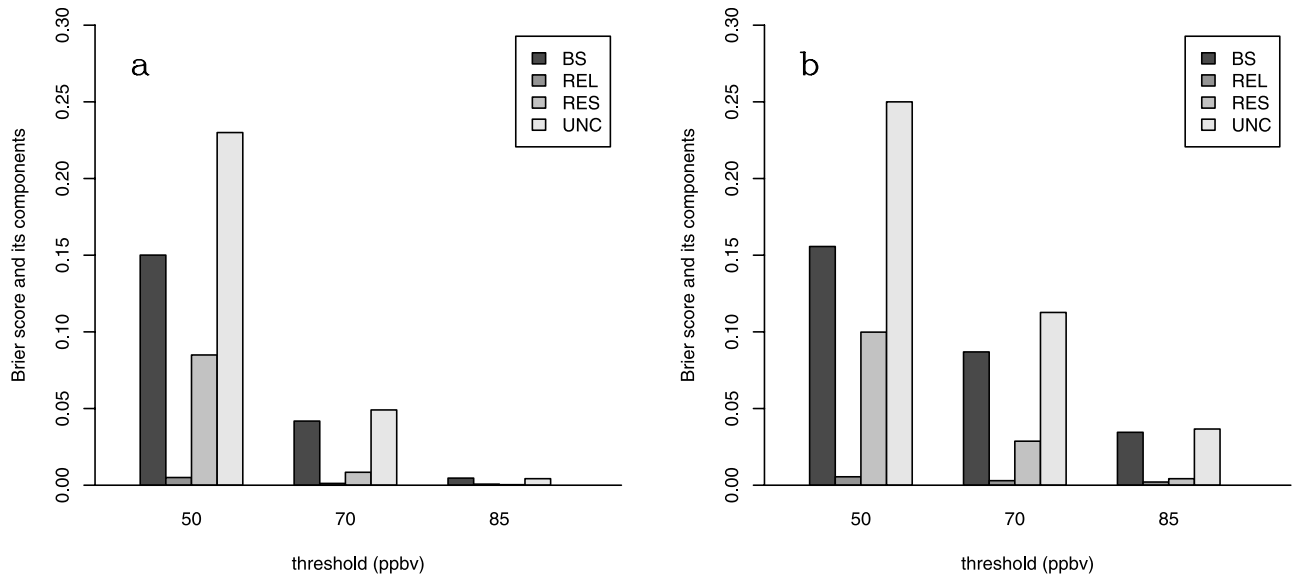

Figure 4. As in Figure 2 but for forecasts calibrated with DLR.

[19] Plots for the corrected ensemble corresponding to those in Figures 2 and 3 are given in Figures 4 and 5. The comparison of *BS*s and their components in Figures 2 and 4 shows positive impact of the DLR on the value of the *BS*, reliability, and, in most cases, also on the resolution. It can be seen by comparing the reliability diagrams in Figures 3 and 5 that with the application of the DLR bias is largely removed while sharpness of the forecasts also improves. The positive effects are the least apparent for the highest threshold and, in the case of the 8-hour average, can possibly be attributed to insufficient sensitivity of DLR to the dependence of the bias on ozone concentrations and the rarity of the events which influences the effectiveness of the DLR.

5. Economic Value of the Forecasts

5.1. Background

[20] Prediction of air quality is of paramount importance for public health and agriculture. Cost-loss analysis certainly has its limitations when human health or loss of life is concerned, but in many instances, can be applied to provide guidance for decision makers by introducing preventive measures restricting human, animal, and plant exposure to pollutants. To assess a potential economic value of surface ozone forecasts, *Richardson*'s [2000, 2003] model will be

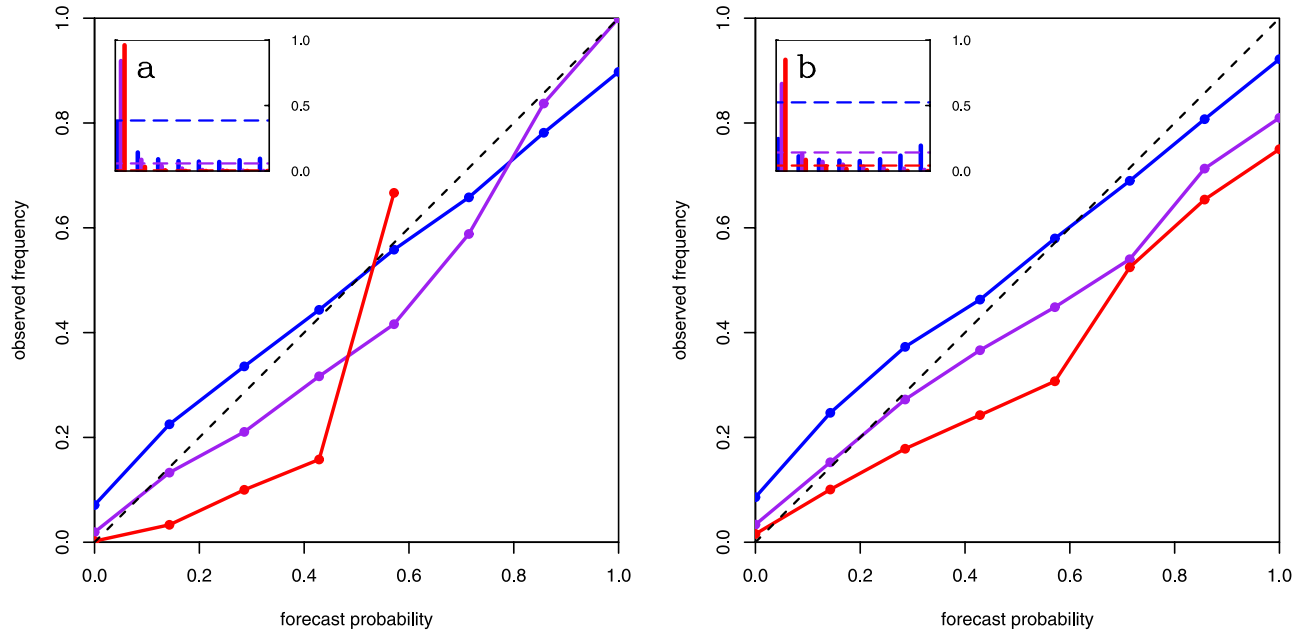

Figure 5. As in Figure 3 but for forecasts calibrated with DLR.

Table 1. Contingency Table

	Observed	Not Observed
Forecast	a (C)	b (C)
Not Forecast	c (L)	d (0)

used. A short description of the model is given below. The reader is referred to the original for full details.

[21] The contingency table shown in Table 1 lists all the possible outcomes of a forecast event or series of events and their counts. For the classification,events need to be binary as defined in section 3. In Table 1, a is a number of hits (observed and forecast), b refers to false alarms (forecast but not observed), c is a number of misses (observed but not forecast), and d is a number of correct rejections (neither observed nor forecast). In the cost-loss analysis model, hits and false alarms incur a cost of taking a preventive action (C), while misses are associated with losses due to the lack of prevention (L). Correct rejections incur no expense. The economic value of forecasts is defined as the reduction in mean expense (defined with respect to climatological forecasts), relative to the reduction (also with respect to climatological forecasts) that would occur if all the forecasts were correct. Using this definition and with assumptions in the contingency table, algebraic manipulation yields the following expression for the economic value:

$$V = \frac{\min(r, o) - F(1-o)r + Ho(1-r) - o}{\min(r, o) - or}, \qquad (3)$$

where o is the climatological frequency of an event, $r = C/L$ is a cost-loss ratio, $H = a/(a + c)$ is a hit rate, and $F = b/(b + d)$ is a false alarm rate. It follows from this formula that for different users of the same forecasts, a potential benefit will vary depending on the application (cost-loss ratio). It can be shown that the economic value given by equation (3) reaches maximum when $r = o$. For deterministic forecasts, the above formula can easily be employed using counts from the contingency table. For the probabilistic forecasts, contingency tables correspond to probability thresholds. An occurrence in a contingency table is counted when probability exceeds a threshold; that is, in the case of an ensemble of m members, if $k + 1$ members predict an event, the count (hit or false alarm) increases in all contingency tables corresponding to probability thresholds from $1/(m + 1)$ to $k/(m + 1)$.

[22] The user of the ensemble forecasts has a choice of taking a preventive action depending on the number of members predicting an event (or forecast probability) and the related cost-loss ratio. It can be shown that for reliable forecasts (i.e., when forecast probability equals observed frequency, see discussion of reliability in the previous section) the optimal decision level occurs when forecast probability equals the cost-loss ratio [*Murphy*, 1977]. In other words, for a reliable (calibrated) forecast the largest benefit is realized if, for a given cost-loss ratio, a preventive action is taken when k members predict an event, i.e., when $p_k = k/(m + 1) \geq r$.

5.2. Application

[23] The application of the above model to assess economic value of the surface ozone forecasts is straightforward. Figure 6 illustrates the method to assess the economic value of ensemble forecasts. Curves are plotted for seven probability thresholds corresponding to seven members of the ensemble by varying the cost-loss ratio. It can be noted that, reasonably, when the expense of prevention is small compared to the expected damage, even small numbers of ensemble members predicting ozone concentration higher than the threshold justify an action. With the increase of the cost-loss ratio, action is required only when a larger number of ensemble members predict the same event. The bold curve which represents the economic value of the ensemble is an envelope of curves plotted for single probabilities. It is apparent that the larger size of an ensemble will generally increase the economic value of forecasts as the cost-loss ratio increment corresponding to the discrete probability thresholds is smaller, so that V approaches $V_{\max}$ for series of $p_k \approx r$. In addition, a larger size of an ensemble will extend the range of forecast probabilities and possibly improve their estimates. No attempt is made here to examine sensitivity of the economic value of the current ensemble to its size since the number of its members is rather small.

[24] In Figure 7, economic values of different deterministic (individual models, ensemble average, and forecasts obtained with DLR) and probabilistic (uncalibrated and calibrated) surface ozone forecasts are shown for 50, 70, and 85 ppbv thresholds for the maximum daily 8-hour and 1-hour averaged concentrations. It can be noted that the general features of the curves do not markedly differ between the 8-hour and 1-hour averages. For the deterministic forecasts, the overforecasting biases of the models (black curves) and the ensemble average (green curves) manifest themselves in the positive economic value for the small cost-loss ratios. It is apparent that the economic value

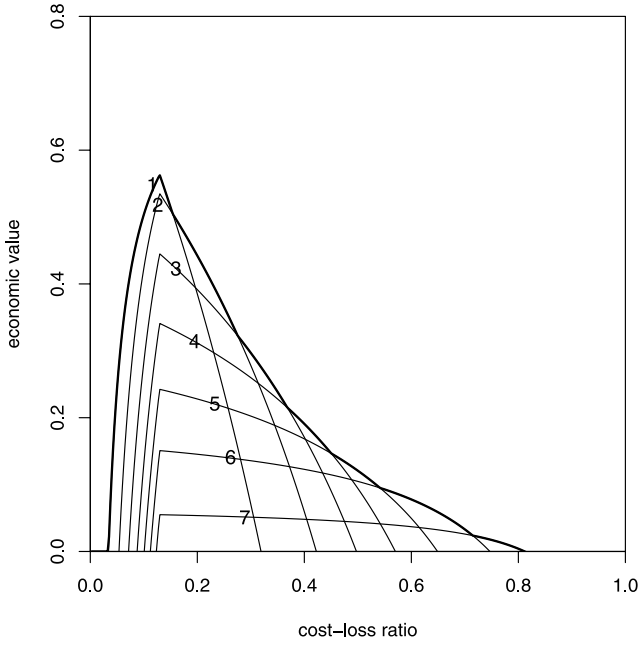

Figure 6. Economic value of probabilistic forecasts with a seven-member ensemble. Thin lines are curves for the probability thresholds equal to 1/8, 2/8, ..., 7/8. Bold line, an envelope of the curves drawn for different thresholds, denotes the ensemble.

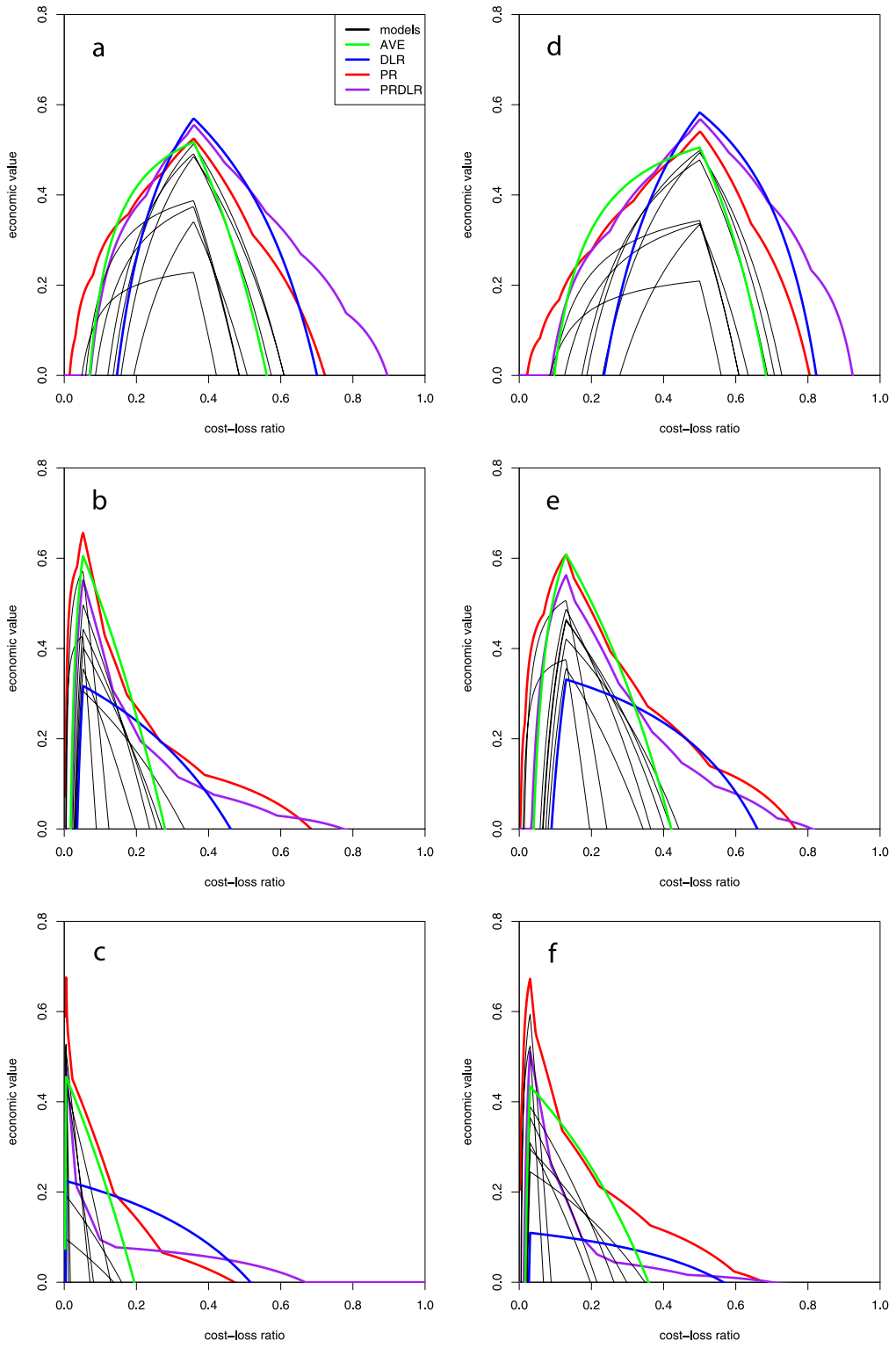

Figure 7. Economic value of deterministic and probabilistic forecasts of maximum daily (a–c) 8-hour and (d–f) 1-hour averaged surface ozone concentration over 50 ppbv (Figures 7a and 7d), 70 ppbv (Figures 7b and 7e), and 85 ppbv (Figures 7c and 7f) thresholds with single models (black), ensemble average (red), ensemble weighted with DLR (blue), uncalibrated probabilistic (red), and probabilistic calibrated with DLR (purple).

of the ensemble average is higher than any of the individual models for all the thresholds (except for the very narrow range of small cost-loss ratios for 85 ppbv). The benefit of DLR (blue curves) is felt with the increase of the cost-loss ratio and is most apparent for 50 ppbv and 70 ppbv thresholds. Uncalibrated probabilistic forecasts (red curves) are superior in terms of economic value to any uncorrected deterministic forecasts except for the middle range of cost-loss ratios for the maximum daily 8-hour average over the 85 ppbv threshold. The economic value of the probabilistic forecasts calibrated with DLR (purple curves) is also generally higher than the economic value of the deterministic forecasts corrected with DLR. Positive effects of the calibration of the probabilistic forecasts with DLR compared to the ensemble average are limited to the 50 ppbv threshold. At 70 ppbv, the threshold economic value of probabilistic forecasts calibrated with DLR is marginally smaller than their uncalibrated counterpart. The economic value of the uncalibrated probabilistic forecasts is clearly superior to their DLR calibrated counterpart for the maximum daily 1-hour average and has a lower/higher value for the small/large cost-loss ratios for the maximum daily 8-hour average. In summary, examination of Figure 7 leads to the following conclusions: (1) the ensemble average provides better economic value than any single model; (2) probabilistic forecasts have generally higher economic value than the deterministic forecasts; and (3) calibration using DLR is generally beneficial for relatively frequent events, and more robust methods of calibration might provide further improvement to forecasts.

6. Summary and Conclusions

[25] Availability of measurements and participation of seven chemical models in the ICARTT/NEAQS field study over the eastern United States and southern Canada in the summer of 2004, provided a unique opportunity to assess advantages of ensemble forecasting of surface ozone concentrations. An assessment of a potential economic value of 24-hour forecasts of the maximum daily 8-hour and 1-hour averaged surface ozone concentrations higher than 50, 70, and 85 ppbv was attempted using the *Richardson*'s [2000, 2003] model. The model defines a measure of economic benefit provided by forecasts, called an economic value, in terms of a reduction in mean expense relative to the reduction obtained with perfect forecasts.

[26] Our results obtained with this economic model demonstrate that the economic value of deterministic forecasts derived for the ensemble of models is superior to the results obtained for the individual models. However, the maximum economic value is achieved by converting forecasts of ensemble members to probabilistic framework in which probabilities are assigned corresponding to the number of members predicting concentrations higher than a threshold. The economic value of probabilistic forecasts is superior to the deterministic forecasts over a wide range of cost-loss ratios and for nearly all the thresholds. The larger economic benefit of probabilistic forecasts is attained through flexibility in the decision on taking a preventive action which is afforded to users by matching forecast probabilities with cost-loss ratios of their specific applications.

[27] Economic value of probabilistic forecasts calibrated with dynamic linear regression [*Pagowski et al.*, 2006] was also attempted. Benefits of such calibrations are limited to lower surface ozone concentrations. It is conjectured that the limited effectiveness of the regression scheme for higher concentration thresholds is a result of generally low ozone in the modeling domain in the summer of 2004. Concentrations of surface ozone higher than the upper two thresholds were rarely measured during the field experiment. Possibly, a more robust method of forecast calibration would yield better results also for the higher concentration thresholds.

[28] The calibration of the ensemble with DLR positively affects Brier scores for all the thresholds through better reliability and resolution (in most cases for the latter). However, smaller values of *BS* do not directly translate to higher economic value of ensemble forecasts as defined and also noted by *Richardson* [2003]. A broader investigation which not only assesses an economic value of probabilistic forecasts but also an amount of information in a wider sense that such forecasts provide, can be envisaged. It is hoped that the results obtained in this paper will encourage more common application of ensembles in air quality in view of potential economic benefits that they carry.

[29] **Acknowledgments.** This research was funded by Early Start Funding from the NOAA/NWS Office of Science and Technology and would not be possible without the participation of the AIRNow program and participating stakeholders. Credit for program and management is given to Paula Davidson (NOAA/NWS/OST), Steve Fine (NOAA/ARL), and Jim Meagher (NOAA/AL). Authors thank participants of the ICARTT/NEAQS field study Veronique Bouchet (Canadian Meteorological Centre), Wanmin Gong (Meteorological Service of Canada), John McHenry (Baron Advanced Meteorological System), Jeff McQueen (NWS/NCEP/NOAA), Richard Moffet (Canadian Meteorological Centre), Steve Peckham (CIRES/NOAA), and Youhua Tang (University of Iowa) who supplied models results. Stu McKeen and Irina Djalalova (both from CIRES/NOAA) helped in processing model and observational data. Stu McKeen also supplied Figure 1 with the map of locations of measurement sites. Dezsö Dévényi commented on the early version of the manuscript. The manuscript was further improved by the critique of reviewers. Susan Carsten and Ann Reiser provided editorial review.

References

Atger, F. (2003), Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration, *Mon. Weather Rev.*, *131*, 1509–1523.

Brier, G. W. (1950), Verification of forecasts expressed in terms of probability, *Mon. Weather Rev.*, *7*, 1–3.

Byun, D. W., and J. K. S. Ching (Eds.) (1999), Science algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System, *EPA-600/R-99/030*, Off. of Res. and Dev., U.S. Environ. Prot. Agency, Washington, D. C.

Carmichael, G. R., et al. (2003), Regional-scale chemical transport modeling in support of the analysis of observations obtained during the TRACE-P experiment, *J. Geophys. Res.*, *108*(D21), 8823, doi:10.1029/2002JD003117.

Delle Monache, L., and R. B. Stull (2003), An ensemble air-quality forecast over western Europe during an ozone episode, *Atmos. Environ.*, *37*, 3469–3474.

Delle Monache, L., X. Deng, Y. Zhou, and R. Stull (2006a), Ozone ensemble forecasts: 1. A new ensemble design, *J. Geophys. Res.*, *111*, D05307, doi:10.1029/2005JD006310.

Delle Monache, L., T. Nipen, X. Deng, Y. Zhou, and R. Stull (2006b), Ozone ensemble forecasts: 2. A Kalman filter predictor bias correction, *J. Geophys. Res.*, *111*, D05308, doi:10.1029/2005JD006311.

Grell, G. A., S. E. Peckham, R. Schmitz, S. A. McKeen, G. Frost, W. Skamarock, and B. Eder (2005), Fully coupled "online" chemistry within the WRF model, *Atmos. Environ.*, *39*, 6957–6975.

Kalnay, E. (2003), *Atmospheric Modeling, Data Assimilation and Predictability*, 341 pp., Cambridge Univ. Press, New York.

Katz, R. W., , and A. H. Murphy (Eds.) (1997), *Economic Value of Weather and Climate Forecasts*, 222 pp., Cambridge Univ. Press, New York.

Kernan, G. L. (1975), The cost-loss decision model and air pollution forecasting, *J. Appl. Meteorol.*, *14*, 8–16.

Mason, I. (1982), A model for assessement of weather forecasts, *Aust. Meteorol. Mag.*, *37*, 75–81.

McHenry, J. N., and C. J. Coats (2003), Improved representation of cloud/actinic flux interaction in multiscale photochemical models, paper presented at 5th Conference on Atmospheric Chemistry: Gases, Aerosols, and Clouds, Am. Meteorol. Soc., Long Beach, Calif.

McKeen, S., et al. (2005), Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004, *J. Geophys. Res.*, *110*, D21307, doi:10.1029/2005JD005858.

Moran, M. D., A. P. Dastoor, S.-L. Gong, W. Gong, and P. A. Makar (1998), Conceptual design for the AES unified regional air quality modelling system (AURAMS), internal report, 100 pp., Air Qual. Res. Branch, Meteorol. Serv. of Can., Toronto, Ont., Canada.

Murphy, A. H. (1973), A new vector partition of the probability score, *J. Appl. Meteorol.*, *4*, 595–600.

Murphy, A. H. (1977), The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation, *Mon. Weather Rev.*, *105*, 803–816.

Mylne, K. (1999), The use of forecast value calculations for optimal decision making using probability forecats, paper presented at 17th Conference on Weather Analysis and Forecasting, Am. Meteorol. Soc., Denver, Colo.

Pagowski, M., et al. (2005), A simple method to improve ensemble-based ozone forecasts, *Geophys. Res. Lett.*, *32*, L07814, doi:10.1029/2004GL022305.

Pagowski, M., et al. (2006), Application of dynamic linear regression to improve the skill of ensemble-based deterministic ozone forecasts, *Atmos. Environ.*, *40*, 3240–3250.

Pudykiewicz, J., A. Kallaur, and P. K. Smolarkiewicz (1997), Semi-Lagrangian modeling of tropospheric ozone, *Tellus, Ser. B*, *49*, 231–258.

Richardson, D. S. (2000), Skill and relative economic value of the ECMWF ensemble prediction system, *Q. J. R. Meteorol. Soc.*, *126*, 649–667.

Richardson, D. S. (2003), Economic value and skill, in *Forecast Verification. A Practitioner's Guide in Atmospheric Science*, edited by I. T. Joliffe and D. R. Stephenson, pp. 164–187, John Wiley, Hoboken, N. J.

Russell, A., and R. Dennis (2000), NARSTO critical review of photochemical models and modeling, *Atmos. Environ.*, *34*, 2283–2324.

Thompson, J. C., and G. W. Brier (1955), The economic utility of weather forecasts, *Mon. Weather Rev.*, *83*, 249–253.

Vautard, R., N. Blond, H. Schmidt, C. Derognat, and M. Beekmann (2001), Multi-model ensemble ozone forecast over Europe: Analysis of uncertainty, paper presented at 26th General Assembly, Eur. Geophys. Soc., Nice, France, 25–30 Mar.

West, M., and J. Harrison (1989), *Bayesian Forecasting and Dynamic Models*, 680 pp., Springer, New York.

Wilks, D. S. (1995), *Statistical Methods in the Atmospheric Sciences. An Introduction*, 467 pp., Elsevier, New York.

Wilks, D. S. (2001), A skill score based on economic value for probability forecasts, *Meteorol. Appl.*, *8*, 209–219.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne (2002), The economic value of ensemble-based weather forecasts, *Bull. Am. Meteorol. Soc.*, *83*, 73–83.

G. A. Grell and M. Pagowski, Global Systems Division, Earth System Research Laboratory, R/GSD1, 325 Broadway, Boulder, CO 80305-3328, USA. (mariusz.pagowski@noaa.gov)